Towards Improved Knowledge Sharing: Assessment of the HL7 Reference Information Model to Support Medical Logic Module Queries

Robert A. Jenders, MD, MS[†]; Walter Sujansky, MD, PhD*;
Carol A. Broverman, PhD[‡]; Michael Chadwick, BA[§]

[†]Department of Medical Informatics, Columbia University
*WiSE Medical Systems, Inc.
[‡]First Databank, Inc.
[§]Chadwick Systems, Inc.

Because clinical databases vary in structure, access methods and vocabulary used to represent data, the Arden Syntax does not define a standard model for querying databases. Consequently, database queries are encoded in ad hoc ways and enclosed in "curly braces" in Medical Logic Modules (MLMs). However, the nonstandard representation of queries impairs sharing of MLMs, an impediment that has come to be known as the "curly braces problem." As a first step in solving this problem, we evaluated the proposed HL7 Reference Information Model (RIM) as a foundation for a standard query model for the Arden Syntax. Specifically, we analyzed the MLM knowledge base at the Columbia-Presbyterian Medical Center and compared the queries in these MLMs to the RIM. We studied 488 queries in 104 MLMs, identifying 674 total query data elements. Laboratory tests accounted for 45.8% of these elements, while demographic and ADT data accounted for 37.6%. Pharmacy orders accounted for 10.5%, medical problems for 4.3% and MLM output messages for 1.6%. We found that the RIM encompasses all but those data elements signifying MLM output (1.6% of the total). We conclude that the majority of queries in the CPMC knowledge base access a relatively small set of data elements and that the RIM encompasses these elements. We propose extensions of this analysis to continue construction of an Arden query model capable of solving the "curly braces problem."

INTRODUCTION

The "Curly Braces Problem"

The Arden Syntax for Medical Logic Modules (MLMs), developed in part at the Columbia-Presbyterian Medical Center (CPMC), has been promoted as an open standard for the procedural representation and sharing of medical knowledge [1]. Implemented using a clinical event monitor at CPMC, MLMs provide over 1000 clinical alerts and many research messages each month [2].

Linking such a knowledge-based system to a clinical database is critical for its effective use [3]. Although the Arden Syntax is a standard for knowledge representation, the Syntax defines only part of a database query. This is explained by the observation that different institutions can vary significantly in the database schema, access methods, query syntax and vocabulary used to store, retrieve and identify clinical data [4]. The site-specific details used to retrieve elements from a database are enclosed within curly braces ("{}") within a query in a MLM. As a result, this problem of site-specificity within an otherwise standard syntax has become known as the "curly braces problem."

A measure of the significance of this problem was demonstrated in a previous analysis of the importance of database links in the CPMC knowledge base in 1993 [3]. In that study of 20 MLMs, the mean number of compiled tokens per data slot was 46% of the total number of tokens. This was the most of any MLM slot measured. Moreover, data queries constituted 65% of the mean execution time of a MLM.

Now, one of the rationales behind the promotion of the Arden Syntax as a standard is the hope that MLMs written at one institution could be shared with another in order to reduce the cost of knowledge engineering [5]. However, the "curly braces problem" interferes with this goal. This was demonstrated in an experiment in which seven CPMC MLMs were shared with LDS Hospital, and the revisions required to permit execution at the other site were analyzed [6]. All 42 data statements in the MLMs required alteration, compared to 20 of the 49 logic statements. Thus, the "curly braces problem"

1091—8280/97/$5.00

compels revision of MLMs when they are used at another site, even though they are written in a standard syntax. This limits sharing and increases the cost of knowledge engineering.

One way to reduce such a cost is the development of a standard query model for the Arden Syntax. Such a model would define a virtual clinical data model against which queries in standard format could be written. Such a model, in keeping with the framework created in the TransFER model [7], would include a standard schema identifying significant data elements and their inter-relationships, a standard query syntax for retrieving such data and a standard vocabulary for identifying them. Automated mapping then could be created between this standard and a local implementation in order to minimize the ad hoc, costly and error-prone translation that must occur now [7]. This would not eliminate the need for local adaptations of MLMs. However, this would increase the ease of sharing knowledge because MLMs would be written in terms of a single data model and query syntax rather than against a myriad of models from different institutions.

The CPMC Architecture

CPMC maintains a central repository of all clinical data. These include laboratory results, inpatient pharmacy orders, demographic data, admission discharge transfer (ADT) data and text reports such as radiology results. CPMC also stores problem lists, outpatient medications and visit notes for a limited number of patients.

As previously described, these data are queried from MLMs using special Data Access Modules (DAMs), providing a relatively uniform interface for queries regardless of the type of database in which these data are stored [2]. Most data are identified using concepts from the CPMC Medical Entities Dictionary (MED), which is the structured, controlled medical vocabulary used at CPMC to identify most data stored in the central repository. These terms are used in MLM query statements [2]. In queries involving ADT or demographic data stored in the hierarchical database (VSAM), segment names instead of MED codes are used to identify data elements; these segments can be matched directly to MED concepts.

Goals of the Analysis

Therefore, as a first step to the solution of the "curly braces problem," we analyzed the current CPMC knowledge base to determine the kinds of queries contained in this production knowledge base. We then compared these query elements to a proposed clinical data model, the Draft HL7 Reference Information Model (RIM) Version 0.8 [8], to assess the adequacy of this standard as a reference schema for clinical data that would satisfy the data needs of typical queries. We conclude by identifying additional work that must be accomplished to complete a standard query model for the Arden Syntax and thus promote easy sharing of MLMs.

METHODS

We reviewed the MLMs stored in the CPMC central repository, dating to December, 1991. From this collection, we eliminated those MLMs created solely for testing or demonstration purposes that were not used in research, hospital administration or clinical production. We then extracted from this collection of MLMs all the Arden "read" statements, eliminating nonfunctional ones (i.e., those that had been "commented out") and those that did not query an external database.

From this list of statements, we extracted the data elements actually retrieved by the queries. We defined a query element to mean a data element that is delineated for retrieval in a query. In the case of queries to the CPMC relational database, query elements are concepts enumerated in the MED. In the case of ADT and demographic data, they are segment identifiers in a hierarchical database, and we mapped these to concepts in the MED for purposes of comparison to relational queries. We then aggregated this list of data elements by term frequency and by synonyms in order to assess the data needs of the CPMC knowledge base.

Proceeding from this aggregation, we then examined the Draft HL7 Reference Information Model Version 0.8 [8] to determine the extent to which this can serve as a standard reference schema of clinical data. Using such a schema, knowledge base developers can formulate database queries in a site-independent way [7].

Although it is not yet an official standard, we chose this model for this purpose because of its broad base of contributors; because of the significant review and comment it has undergone and will undergo within the HL7 community; and because of our estimation of its probability of success in becoming an accepted if not official clinical data model. Another candidate model is the ASTM E1384-96 standard for content

and structure of the computer-based patient record. However, we felt that this is less widely accepted than the HL7 standard and hence less likely to be maintained. Another possible model is the CPMC MED. However, this is not accepted as a standard and contains many site-specific concepts.

The RIM is divided into subject areas relevant to some aspect of health care data, such as "health care stakeholders" and "patient encounters." These are further decomposed into 126 classes, each of which has relationships to other classes and is further described by attributes. We compared each query element in the aggregate list from the CPMC MLMs to the textual descriptions of the RIM classes and their attributes. Our objective was to assess whether each element was definitely encompassed, possibly encompassed or not encompassed by some RIM class. A "definitely encompassed" element is one that constitutes a subtype or instance of the corresponding RIM class, that is, it could be stored in a relational table or hierarchical collection represented by that class.

RESULTS

We isolated 104 qualifying MLMs from the CPMC knowledge base. From this collection we extracted 488 queries to the CPMC central repository. These queries contained a total of 674 query elements and 179 distinct query elements. This yields a mean of 1.4 data elements per query and 4.7 queries per MLM.

We extended this analysis by aggregating individual query elements that are synonymous. For example, multiple synonyms of "blood glucose measurement" appear separately in the simple frequency listing but may be aggregated into a single semantic concept. The fifteen most prevalent aggregate concepts appear in Table 1. Of these most prevalent concepts, five are laboratory tests, accounting for 32.5% of data elements in the query set; four are ADT elements (nursing unit, case type, discharge date and physician name), accounting for 14.1%; and two are demographic elements (patient name and birthdate), accounting for 12%; Patient problems (diagnosis code and text) account for 3.8%, while pharmacy queries (aminoglycoside antibiotics) account for 2.4% and decision support system output (MLM message) for 1.6%. All told, the fifteen most prevalent data concepts account for nearly 2/3 (66.4%) of all data elements queried.

Table 1. Fifteen most frequent query elements, aggregated by synonymy. Frequency denotes the number of times the element appeared in the CPMC query set. Percent denotes the percentage of all query elements represented by an element.

Query Element	Frequency	Percent
Blood Glucose	106	15.7
Blood Creatinine	48	7.1
Patient Name	44	6.5
Nursing Unit	42	6.2
Blood Potassium	38	5.6
Birthdate	37	5.5
Case Type	23	3.4
Blood pH	21	3.1
Discharge Date	16	2.4
Aminoglycoside Antibiotics	16	2.4
Physician Name	14	2.1
Diagnosis Code	13	1.9
Diagnosis Text	13	1.9
MLM Output	11	1.6
Blood Sodium	7	1.0
TOTAL	**449**	**66.4**

Note that some query elements, such as "patient name" are explicit attributes in some data models (e.g., RIM), while other elements, such as "blood glucose," are vocabulary terms that might be instantiations of classes (such as "clinical observation" or "service event" in the case of blood glucose) in a data model. Thus, the CPMC query set mixes vocabulary terms and data model elements.

We then further abstracted the original set of MLM query elements into higher-level subject areas corresponding to the kind of data being requested, e.g., ADT, laboratory results, pharmacy orders and demographic data. The frequency of these areas among the query elements in the set is detailed in Table 2.

We subsequently compared the query elements identified in our query set, along with abstractions of these elements, to the relevant subject areas and classes in the HL7 RIM. We found that all of the CPMC laboratory result query elements were definitely encompassed by the RIM class "clinical_observation." Query elements describing problems all were definitely encompassed by the RIM class "patient_diagnosis". Pharmacy order data elements all were definitely encompassed by the RIM

class "patient_service_order," which is a generalization of the class "pharmacy_service_order." Data elements concerning ADT were definitely encompassed by several RIM classes, including "adt_order" and "episode." In addition, demographic data were definitely encompassed by assorted RIM classes.

As a simplified example, a typical CPMC MLM statement using the most commonly accessed query element (blood glucose) is "read last {'dam' = "PDQRES2", 'constraints' = "C****";; '32308'}", where '32308' is the MED code for "intravascular glucose test". Using the RIM as the standard data model and presupposing a SQL-like query syntax, this could be rendered instead as "SELECT last observation_value_text FROM clinical_observation WHERE abbreviation_name ISA "blood sugar".

Table 2. High-level subject areas of query elements. Frequency and percent are defined in Table 1. Percent does not add to 100 because of rounding.

Subject Areas	Frequency	Percent
Laboratory Result	309	45.8
ADT	150	22.2
Demographic Data	104	15.4
Pharmacy Orders	71	10.5
Problems	29	4.3
MLM Output	11	1.6

The only aggregate of query elements that was *possibly* encompassed (rather than definitely encompassed) by a RIM class was MLM Output. MLM Output elements probably correspond to the RIM class "clinical_observation_report," which is defined by HL7 as "a report of the information or record secured by an act or instance of viewing or noting a fact or occurrence for some health related purpose." [8] Although this seems to be the most closely matching class in the RIM, there is not an unequivocal correspondence. However, in the CPMC knowledge base, MLM Output elements constitute only 1.6% of all query elements.

DISCUSSION

The problem of having to revise a rule in a computer-based expert system in order that it may work in a site other than the one for which it was originally composed is not peculiar to the Arden Syntax. Any procedural chunk of knowledge, no matter the knowledge representation scheme, will require site-specific adaptation to local architecture, database schema and vocabulary. Arden is especially associated with this challenge, however, because it is a standard for procedural medical knowledge representation whose original design objective was knowledge sharing.

We have proposed a reference data model as the first step toward solving the "curly braces problem" of the Arden Syntax. Our evaluation of the current CPMC knowledge base of Arden MLMs demonstrates that most query elements that appear in these MLMs are encompassed by relatively few clinical identifiers. Moreover, many of these query elements are encompassed by classes of the HL7 Reference Information Model. Therefore, we believe that the RIM is a good starting point in the creation of a standard Arden query model.

One weakness in this assertion is the fact that the RIM is only a draft model and may undergo significant revision before being accepted as a standard.

Another weakness of the current analysis may be the relatively limited size of the query set, which may bias the kinds of queries found therein. However, the CPMC knowledge base has existed for over five years. Accordingly, we believe that its queries, even if not entirely representative of those at other institutions, likely represent a sizable portion of the data elements that any working knowledge base will require in a production environment. Thus, we believe that any reference data model that can encompass the query elements in this knowledge base likely will be adequate at other sites.

This assertion in turn is somewhat limited by the data available in the CPMC repository for access by MLMs. For example, the CPMC architecture does not currently include an electronic order-entry system. Thus, many rules which might be written to analyze physician orders cannot function at CPMC because of the absence of these data. However, we note that the RIM includes many classes related to orders, and therefore it is likely to encompass adequately query elements related to order-entry MLMs.

FUTURE WORK

Defining a reference data model such as the HL7 RIM is only the first step in the creation of a uniform query model that can be incorporated in the Arden Syntax. Other essential parts of the model include a

standard query syntax and a standard vocabulary. Likely the model will require several different vocabularies to represent different kinds of data.

By further studying the CPMC and other knowledge bases, we hope to extend the current analysis to identify and evaluate existing query syntax and vocabulary models. In doing so, we hope to provide a complete, uniform query model that can be incorporated into the Arden standard. Using such a model, MLM writers at any institution can compose queries against the virtual clinical database defined by the model, rather than against their local idiosyncratic clinical databases. In turn, this will reduce the challenge of incorporating MLMs written at various sites into a particular site's knowledge base because any site need only accomplish one mapping to its local implementation--rather than many different mappings between the disparate data models at other sites and its own data model. This facilitation will promote knowledge sharing and help advance a significant goal of the Arden Syntax.

SUMMARY

The knowledge base of Arden Syntax MLMs at the CPMC provides the basis for a robust decision support system. In an examination of 104 MLMs, we identified 488 queries with 674 total and 179 distinct data elements. Nearly 2/3 of these query elements can be aggregated into only 15 concepts. Laboratory, ADT and demographic data account for over 3/4 of the data elements in this query set. The HL7 Reference Information Model encompasses nearly all of the query elements in the set. Therefore, this model can serve as the starting point for the construction of a uniform Arden query model. Additional work to define a standard query syntax and vocabularies must be accomplished in order to realize this goal.

Acknowledgments

We thank Balendu Dasgupta for his programming assistance in the completion of this analysis. We are indebted to George Hripcsak, Stephen Johnson, James J. Cimino and Paul Clayton, whose efforts were instrumental in the creation of the CPMC decision support system. This work was supported in part by continuing generous assistance from The Presbyterian Hospital in the City of New York and the Center for Advanced Technology at Columbia University.

References

1. Hripcsak G, Clayton PD, Pryor TA, Haug P, Wigertz OB, Van der lei J. The Arden Syntax for medical logic modules. In Miller RA, editor. Proc. of the Fourteenth Annual Symposium on Computer Applications in Medical Care. New York: IEEE Computer Press, 1990; 200-204.

2. Jenders RA, Hripcsak G, Sideli RV, DuMouchel W, Zhang H, Cimino JJ, Johnson SB, Sherman EH, Clayton PD. Medical decision support: experience with implementing the Arden Syntax at the Columbia-Presbyterian Medical Center. In Gardner RM, ed. Proc. of the Nineteenth Annual Symposium on Computer Applications in Medical Care. Philadelphia: Hanley & Belfus, 1995; 169-173.

3. Hripcsak G, Johnson SB, Clayton PD. Desperately seeking data: knowledge base-database links. In Safran C, ed. Proc of the Seventeenth Annual Symposium on Computer Applications in Medical Care. New York: McGraw-Hill, 1993; 639-43.

4. Scherpbier HJ, Klein SR, Perreault L, Jenders RA. Aspects of knowledge sharing using the Arden Syntax. Proc. of the Annual Health Information and Management Systems Society Conference, 1996; 2:111-122.

5. Hripcsak G, Ludeman P, Pryor TA, Wigertz O, Clayton PD. Rationale for the Arden Syntax. Comput Biomed Res 1994; 27:291-324.

6. Pryor TA, Hripcsak G. Sharing MLM's: An experiment between Columbia-Presbyterian and LDS Hospital. In Safran C, ed. Proc of the Seventeenth Annual Symposium on Computer Applications in Medical Care. New York: McGraw-Hill, 1993; 399-404.

7. Sujansky W, Altman R. Towards a standard query model for sharing decision-support applications. In Ozbolt JG, ed. Proc of the Eighteenth Annual Symposium on Computer Applications in Medical Care. Philadelphia: Hawley and Belfus, 1994; 325-31.

8. Health Level Seven. Draft HL7 Reference Information Model. Version V 0.08, 1997.